Group Method of Data Handling: How does it measure up?

Poorani Selvaraj, Gary R. Weckman
Department of Industrial and Systems Engineering
Ohio University
Athens, Ohio USA
Email: ps365012@ohio.edu, weckmang@ohio.edu

Andrew P. Snow
School of Information & Telecommunication Systems
Ohio University
Athens, Ohio USA
Email: snowa@ohio.edu

***Abstract*— Prediction is the method of determining future values based on the patterns deduced from a data set. This research compares various data mining techniques—namely, multiple regression analysis in statistics, Artificial Neural Networks (ANN), and the Group Method of Data Handling (GMDH), including both with and without feature selection. Currently, the literature suggests that GMDH, an inductive learning algorithm, is an excellent tool for prediction. GMDH builds gradually more complex models that are evaluated via a set of multi-input, single-output data pairs. ANNs are inspired by the complex learning that happens in the closely interconnected sets of neurons in the human brain and are also considered an excellent tool for prediction. This article is the beginning of a more detailed research project to investigate how well GMDH performs in comparison to other data mining tools.**

Keywords- Group Method of Data Handling; Artificial Neural Networks; Prediction; Statistics; Feature Selection.

I. INTRODUCTION

In this day and age, a large amount of profitable information is being processed from the myriad of data that is being collected. Why is this information significant, who uses this information, and for what purposes is it being used? To be able to answer these questions, we will begin asking the very basic question: What is data? Data is information of any nature that, when quantized, can be meaningfully disseminated into useful knowledge [1].

There has been no discrimination in regards to the type of industries from which databases emerge. Disparate industries have understood the need to exploit the knowledge that can be extracted by scouring through these large repositories of data. Knowledge is a term that is commonly associated with data [1]. For example, the itemization of a grocery bill, along with its corresponding loyalty card number, is generally considered data. This data may be used by data scientists to estimate the number of people in the household, the age group, and so on. This is known as knowledge, which is extracted based upon data obtained from the bill. However, the biggest challenge that we face is disentangling the useful knowledge that is desegregated from all the noise, but that is also collected as part of the data in a repository[1][2]. Data mining helps in addressing this data overload problem that we face at a time when the world is progressing towards an era of digital information.

Data mining is the art of extracting understandable patterns that may be concealed within a vast repository of data and identifying potentially useful information that can be used to our advantage [3]. The choice of the proper data mining technique among those that are usually used depends on the kind of information we wish to extract from the data. Depending upon the type of knowledge we wish to gain, data mining usually involves six common classes of applications.

The process of detecting interesting relationships between attributes is known as association. This type of learning commonly explores large spaces of potential patterns and chooses those that may be of interest to the user, which are generally specified using a constraint [4]. This application of data mining is most commonly used in the business world, where they base most of their micro- and macro-business decisions off of these patterns

The use of a model to fit data into pre-categorized discrete classes is known as classification [5]. According to Zhang and Zhou, classification is the process of identifying common features that describe and distinguish common classes [6]. E.W.T. Ngai et al. suggest that the most commonly used techniques for the classification of data include neural networks, the naïve Bayes technique, decision trees, and support vector machines [5].

According to E.W.T. Ngai et al, dividing objects that are similar to each other in order to form groups that are conceptually meaningful—and, at the same time, very dissimilar to those from the other group--is called clustering [5]. Zhang and Zhou explain this as maximizing "intra-class" similarity and minimizing "inter-class" similarity [6]. The clusters are usually mutually exclusive and exhaustive. For a more complex and in-depth representation of the data, the model may be chosen to represent a hierarchical or overlapping category [7].

For estimation, we find an approximate value of a target or dependent variable using a set of predictor or independent variables [8]. Regression analysis is used for this purpose. Regression analysis is the generation of an equation that best represents the relationship between a continuous dependent output variable and one or more independent input variables [5]. It is mostly a statistics-based methodology that is used for estimation and forecasting. Based on the number of independent or predictor variables, a simple linear regression

 ISBN: 978-1-61208-457-2

or a multiple linear regression may be performed. A well-fit model is one in which a strong correlation exists between the two variables.

Prediction, as the name suggests, is the method of determining future values based upon the patterns deduced from the data set. It is very similar to classification and estimation--however, a very fine line exists between them. According to E.W.T. Ngai et al, the most common techniques used for prediction analysis are neural networks and logistic model predictions [4].

In Section 2 and 3, we explain the techniques we use for analysis. In Section 4 and 5, we discuss the methodology and the database used. In Section 6, we discuss the results and the overall conclusion is stated.

II. ARTIFICIAL NEURAL NETWORKS

ANNs are inspired by the complex learning that happens in the closely interconnected sets of neurons in the human brain [8]. An analogy between a neural network and a human brain can be drawn in the following manner [9], as illustrated in Figure 1.

An ANN always acquires its knowledge through learning, similar to a human brain, which is constantly learning through experiences. The ANN's knowledge is stored in its neurons, using weights associated in its inter-neuron connections, which are similar to the synaptic weights that we have in the neurons in a human brain.

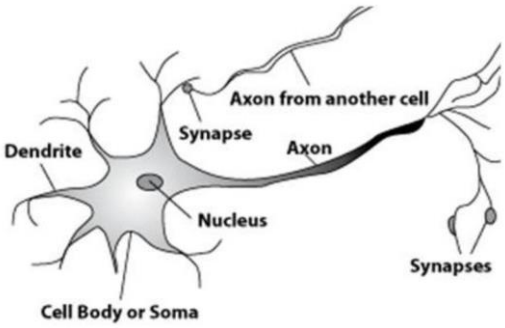

Figure 1. Biological neuron [9].

A generic structure of a neural network model can be explained as a mathematical equivalent, which is shown in Figure 2.

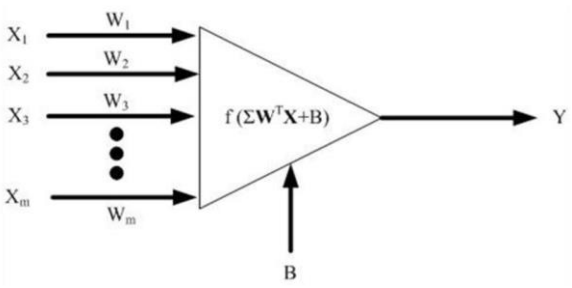

Figure 2. Arificial Perceptron [9]

Figure 2 displays a mathematical approximation of the biological model in Figure 1; for the mathematical model, a linear combination of weights and input values are passed through activation functions that process the data. For example, the perceptron in Figure 2 has independent inputs (*X1*, *X2*, *X3*, and *Xm*), connection weights (*W1*, *W2*, *W3*, and Wm), a bias (*B*), and a dependent variable (*Y*). During this operation, the perceptron computes a weighted sum as each input and bias passes through the neuron. Weights represent the strength of the association between independent and dependent variables, and can be positive, negative, or zero; the weighted sum is then processed by the neuron's activation function and sent through the rest of the network.

The mathematical expression of the perceptron neuron shown above is:

$$y_k = \varphi(\sum w_{km} * x_m + b_k) \quad (1)$$

Where w_{km} =connection weight of source neuron *m* to target neuron *k*; $b_k = \text{bias}$; $y_k = \text{output}$; $\varphi = \text{transfer function}$. One of the major advantages of ANN is their robust nature that allows them to represent and learn both linearly- and non-linearly-related data with ease [9].

A multilayer perceptron (MLP) is a modification of the simple linear perceptron, and can easily model highly nonlinear data of input vectors to accurately represent the new unseen data [11][12]. A simple MLP structure is illustrated in Figure 8 (end of article), and demonstrates how the perceptron fits into an ANN.

The architecture of an MLP can usually vary, but in general, it has input neurons which provide the input parameters to the network. There may be one or more hidden layers, which are made up of nodes connected to every other node in the prior and subsequent layers. The output signals from these nodes are the sum of the input signals to the node, which may be modified by an activation function or transfer function, which is generally non-linear. We make use of nonlinear functions to produce the outputs, which are then scaled using the weights associated with the nodes in order to be fed forward as the input to every node in the next layer of the network. This method of information processing, which is directed in the forward path of a network, makes the MLP a feed forward network [11].

By constant training, the MLP network has the ability to learn and to accurately model the input-output relationship. A set of training data with input vectors and target output vectors is used for the purposes of training. While training, the model constantly adjusts the weights of the nodes until the magnitude of the error of the MLP network is minimized. This is performed by constantly monitoring the error that arises as a difference between the predicted and the actual output, and adjusting the weights accordingly [11]. Thus, the MLP uses a supervised learning technique. The most commonly used algorithm for this purpose is the back-propagation algorithm. This training of the MLP is stopped when the performance of the network can accurately predict the target variable, which is compared to the testing data set.

 ISBN: 978-1-61208-457-2

III. GROUP METHOD OF DATA HANDLING

Ivakhnenko in 1966 introduced the concept of GMDH as an inductive learning algorithm [13][14]. According to Ravisankar et al, GMDH builds gradually more complex models that are evaluated on a set of multi-input, single-output data pairs [15][16][17]. Dipti suggests that the complexities involved in other neural networks—such as determining the most important input variables, the number of hidden layers, and the neurons--are all circumvented by GMDH [14][18]. The need for prior knowledge to build models is eliminated by GMDH, and hence, it is a self-organizing feed forward neural network.

The neurons use competitive learning, as opposed to back propagation error correction. The competitive learning is based upon the way that the neurons compete with each other in order to respond to the input neurons. The overall methodology includes the following steps:

A data set with n observations for regression analysis is collected, with m independent variables x_i and a dependent variable y_j; i=1, 2, 3….m; j=1, 2, 3….n

Step 1: The data set is divided into training and checking sets.

Step 2: A regression equation for each pair of independent variables is computed as follows :

$$y = A + Bx_i + Cx_j + Dx_i^2 + Ex_j^2 + Fx_{ij} \quad (2)$$

leaving us with $\binom{m}{2} = m(m-1)/2$ sets of regression polynomials, each made up of pairs of independent variables. We now have higher order variables predicting the output, as opposed to the original m variables $x_1, x_2, x_3 \ldots x_m$.

Step 3: Each of these regression surfaces will then be evaluated at all n data points. For example, a regression equation of the first two independent variables x_1 and x_2 is generated and then evaluated against all n data points as (x_{11}, x_{12}), (x_{21}, x_{22}), (x_{31}, x_{32})…….(x_{n1}, x_{n2}). This is now stored as a new variable Z_1. The remaining variables are computed in a similar manner. It is common knowledge that these new variables predict the output better than the original independent variables $x_1, x_2 \ldots x_m$.

Step 4: We now choose survivors, or those Z variables that best represent the output variable y by evaluating it against the checking set. The survivors are calculated by estimating the regularity criterion, which is usually the mean squared error (r_{min}), and arranging the values in increasing order of the regularity criterion for each Z variable. Based upon a pre-determined value R, those values of Z whose mean squared error is less than that of R are chosen as the survivors to replace the corresponding values of x.

The whole process is repeated, and we now have a regression polynomial of order four. In this way, the model builds gradually, complicating polynomial models of increasing order until the r_{min} value of one model is no longer less than the previous model. This now implies that the r_{min} value has reached its minimum and we can stop the process.

Feature selection is the process of using a smaller set of features, predictors, or independent variables to describe a sample in the measurement space. According to Guyon and Elisseeff, there are a number of potential benefits by this method of feature selection [19]:

- Helps in the visualization and better understanding of the data
- Feature selection enables reduced storage requirements
- Helps reduce the time to train the network
- Reduces the dimensionality of the network and improves the prediction performance.

For this research, GMDH is used to select the most significant features, depending upon their ability to have the best accuracy in their test data set. Ten-fold cross validation is performed, and the features which have the most frequency of occurrence in all of the 10 folds were selected based upon the percentages that were obtained from the GMDH Shell software.

IV. METHODOLOGY

The flow chart below gives an outline of the methodology that has been followed for analyzing data for the purposes of comparing various data mining techniques.Namely, this includes multiple regression analyses in statistics, MLP, and GMDH, including those both with and without feature selection. See Figure 9.

To compute various statistical data sets usually for six sigma initiatives--Minitab is used. It is a very versatile and effective tool [20]. It provides tools and options for both basic and advanced data analysis.

To compute neural network models, NeuroSolutions [21], an easy-to-use neural network software package for Windows, is used. It provides an easy-to-use Excel interface and a user-friendly intuitive wizard with an icon-based network design interface, which are used to implement advanced artificial intelligence and learning algorithms [9].

GMDH Shell is used for the purpose of accurately forecasting time series, build classification, and regression models. It is also neural network-based software that allows for a full spectrum of parametric customization. However, the differentiating factor is that it is very fast, since it implements advanced parallel processing and has highly optimized the core algorithms. It can be used for any task from data sciences and financial analysis to inventory forecasting, demand forecasting, load forecasting, demand and sales forecasting, and stock market prediction [22][23].

In this broad range of knowledge discovery applications, the main idea is to train a subset of the population with known labels, and then make the predictions for a test subset with unknown labels [24]. When the learning algorithm is trained using only the training set, the algorithm looks for patterns in the training set which are depictive of the correlations that exist between the features and output of the data set. However, it is important to note that these patterns may be specific to only the training data set, i.e., they are valid only in the training set, and are not actually true for the

general population of the database [24]. This will cause the algorithm to have a higher accuracy rate in the training set. It is not uncommon for the learning algorithm to become one hundred percent accurate with issues of over fitting. The tailoring of the algorithm to match the patterns that may be unique to only the training set--which might be due to randomness in how the training data was selected from the population--is called over fitting. In order to avoid these issues, a validation subset which is mutually exclusive of the training set is used to fine-tune the model.

- Training set: To fit the parameters, i.e., weights
- Validation set: To tune the parameters, i.e., architecture
- Test set: To assess the performance. i.e., generalization and predictive power

For the purpose of our research, we have divided 60 percent of the data set as a training set, 20 percent as cross validation, and 20 percent as a test set. The data set was initially randomized to avoid any discrepancies that may have occurred while experimenting.

V. CASE STUDY

This case study is based on data collected by the National Oceanic and Atmospheric Administration (NOAA) study addressing the impacts of mussel recruitment on the bay's water quality via provided hydrological data (April through November, 1991–1996) [25]. The database includes the following variables: temperature (TEMP), Secchi depth (SECCHI), light attenuation (Kd), total suspended solids (TSS), TP, soluble reactive phosphorus (PO4)3-P), nitratenitrogen (NO3)-N), ammonia-nitrogen (NH4 +-N), silica (SiO2), particulate silica (PSiO2), particulate organic carbon (POC), chloride (CL), total photosynthetic radiation (TotPAR), visibility (VISIB), ambient temperature (TEMPAmb), and wind speed (WNDSpd). The dataset consists of 251 records [25].

Saginaw Bay (See Figure 3) is part of Lake Huron. Prior to the 1980s, excessive nutrient loading altered the water quality and resulted in expansive cyanobacterial blooms during the summer months. Bloom intensity and frequency decreased in the mid-1980s following the initiation of nutrient-abatement programs. Invasion of mussels occurred during the early 1990s, with larvae and adult mussels first being observed in 1991. During 1994–1996, their growth stabilized and populations became established (Nalepa et al. 1995). Coincident with mussel occurrence, blooms of toxic Microcystis reappeared during late summer months, and have remained annually recurrent throughout the bay [25].

Saginaw Bay is divided into two regions. Water quality in the shallow ''inner bay'' largely is influenced by nutrient-laden inflows of the Saginaw River. The river drains from agricultural, industrial, and urban areas. The mean circulation is weak, and water exchanges between the inner and outer bays occur along the northern shorelines.

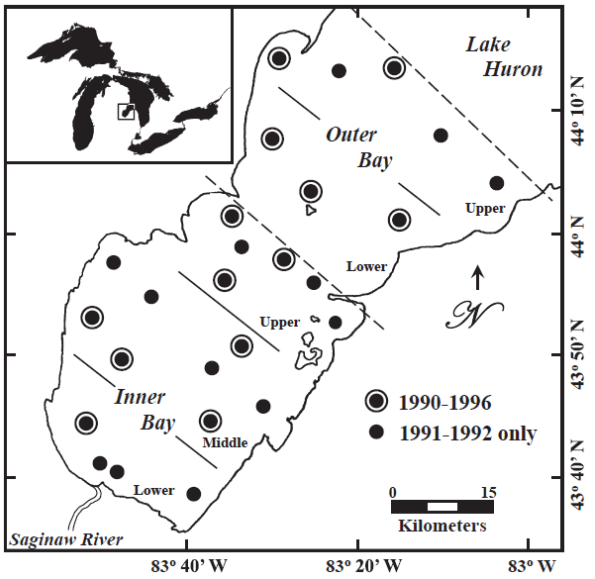

Figure 3. Location of sampling stations within Saginaw Bay [25].

VI. RESULTS AND CONCLUSIONS

Many models were constructed, trained, and tested, as described earlier. The performance metric used is R^2--which is commonly called the coefficient of determination or the coefficient of multiple determination for multiple regression. The results of this analysis are summarized in Table 1:

TABLE 1: COMPARISON OF DATA MINING TECHNIQUES

	Train	Cross Validation	Test
Statistics	0.843		0.685
GMDH	0.929	0.912	0.890
MLP	0.945	0.951	**0.927**
GMDH - FS	0.924	0.912	0.850
MLP - FS	0.977	0.949	**0.922**

As noted in the table, the MLP outperformed both the GMDH and basic statistics, whereas GMDH outperformed just statistics. In this case, the feature selection (FS) did not seem to have a benefit in MLP or GMDH, except in terms of reducing the number of attributes in the model. Figures 4 through 7 illustrate how well the model actually predicted the output values versus actual values. The perfect model would fit the 45-degree diagonal line. In reviewing the figure, it can be seen that the MLP model has the least variation from the diagonal line.

This research only looked at one database, so the conclusions are very limited in scope. Additional research is being performed, and the results of a more extensive study will be published in the near future.

 ISBN: 978-1-61208-457-2

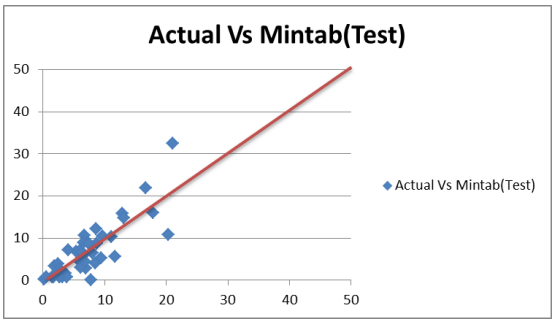

Figure 4. Comparison of Actual versus Minitab.

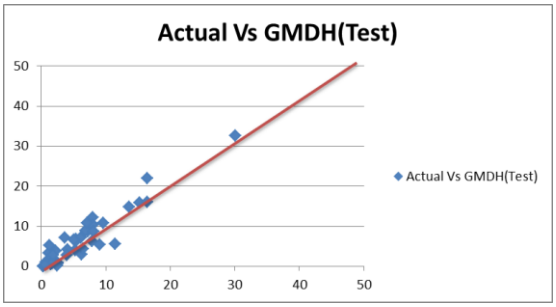

Figure 5. Comparison of Actual versus GMDH.

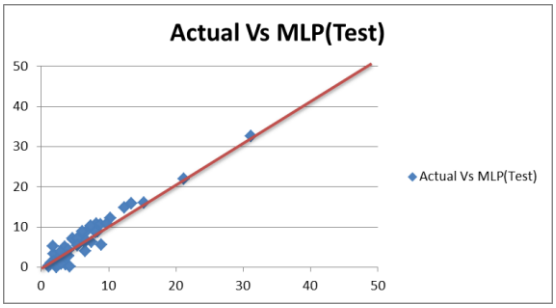

Figure 6. Comparison of Actual versus MLP.

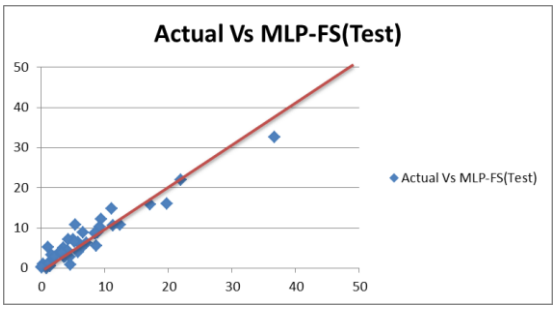

Figure 7. Comparison of Actual versus MLP with feature selection.

REFERENCES

[1] G. Florin, Data mining: concepts, models and techniques. Berlin, Heilelberg: Springer-Verlag, 2011.

[2] L. A. Kurgan and P. Musilek, "A survey of Knowledge Discovery and Data Mining process models," Knowl. Eng. Rev., vol. 21, no. 01, Mar. 2006, pp.1-24.

[3] U. M. Fayyad, P. Stolorz, "Data mining and KDD:Promise and challenges," Future Gener. Comput. Syst., vol. 13, no. 2-3, Nov. 2007, pp. 99-115.

[4] G. I. Webb, "Discovering Significant Patterns," Mach. Learn., vol. 68, no. 1, May 2007, pp. 1–33.

[5] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," Decis. Support Syst., vol. 50, no. 3, Feb. 2011, pp. 559–569.

[6] D. Zhang and L. Zhou, "Discovering Golden Nuggets: Data Mining in Financial Application," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 34, no. 4, Nov. 2004, pp. 513–522.

[7] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," AI Mag., vol. 17, no. 3, 1996, pp. 37-54.

[8] D. T. Larose and C. D. Larose, Wiley Series on Methods and Applications in Data Mining: Data Mining and Predictive Analysis, Second. John Wiley & Sons, Inc, 2015.

[9] W. A. Young II, W. S. Holland, and G. R. Weckman, "Determining Hall of Fame Status for Major League Baseball Using an Artificial Neural Network," Journal of Quantitative Analysis in Sports, vol. 4 : iss. 4, no. 4 Oct. 2008, pp. 1131-1135.

[10] D. J. Fonseca, D. O. Navaresse, and G. P. Moynihan, "Simulation metamodeling through artificial neural networks," Eng. Appl. Artif. Intell., vol. 16, no. 3, Apr. 2003, pp. 177–183.

[11] M. W. Gardner and S. R. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," Atmos. Environ., vol. 32, no. 14, Aug. 1998, pp. 2627–2636.

[12] T. Isokawa, H. Nishimura, and N. Matsui, "Quaternionic Multilayer Perceptron with Local Analyticity," Information, vol. 3, no. 4, Nov. 2012, pp. 756–770.

[13] S. J. Farlow, Self-Organizing Methods in Modeling: GMDH Type Algorithms. CRC Press, 1984.

[14] D. Srinivasan, "Energy demand prediction using GMDH networks," Neurocomputing, vol. 72, no. 1–3, Dec. 2008, pp. 625–629.

[15] P. Ravisankar and V. Ravi, "Financial distress prediction in banks using Group Method of Data Handling neural network, counter propagation neural network and fuzzy ARTMAP," Knowl.-Based Syst., vol. 23, no. 8, Dec. 2010, pp. 823–831.

[16] P. Ravisankar, V. Ravi, G. Raghava Rao, and I. Bose, "Detection of financial statement fraud and feature selection using data mining techniques," Decis. Support Syst., vol. 50, no. 2, Jan. 2011, pp. 491–500.

[17] M. Mottaghitalab, A. Faridi, H. Darmani-Kuhi, J. France, and H. Ahmadi, "Predicting caloric and feed efficiency in turkeys using the group method of data handling-type neural networks," Poult. Sci., vol. 89, no. 6, Jun. 2010, pp. 1325–1331.

[18] G. C. Onwubolu, "Design of hybrid differential evolution and group method of data handling networks for modeling and prediction," Inf. Sci., vol. 178, no. 18, Sep. 2008, pp. 3616–3634.

[19] I. Guyon and A. Elisseeff, "Empirical Inference for Machine Learning and Perception Department", The Journal of Machine Learning Research, vol 3, Mar. 2003, pp. 1157-1182.

[20] "Review of Minitab 15 Software Program for Use in Six Sigma," Brighthub Project Management. [Online]. Available: http://www.brighthubpm.com/software-

 ISBN: 978-1-61208-457-2

reviews-tips/33580-review-of-minitab-fifteen-for-six-sigma/. [Accessed: 15-Oct-2015].

[21] "Neurosolutions7". [Online]. Available: http://www.neurosolutions.com/neurosolutions/help/. [Accessed: 29-Sep-2015].

[22] "GMDH Shell | Binary Today." [Online]. Available: http://www.binarytoday.com/gmdh-shell/. [Accessed: 21-Oct-2015].

[23] "Data Mining Solution for Business", GMDH Shell Forecasting and data Mining Software." [Online]. Available: https://www.gmdhshell.com/data-mining-software. [Accessed: 18-Sep-2015].

[24] C. Elkan, "Evaluating classifiers," Univ. San Diego Calif. Retrieved 01-11-2012 Httpcseweb Ucsd Edu~ Elkan B, vol. 250, 2012.

[25] D. F. Millie et al, "An 'enviro-informatic' assessment of Saginaw Bay (Lake Huron USA) phytoplankton: characterization and modeling of Microcystis (Cyanophyta)," Journal of Phycology, vol. 47, no. 04, Aug. 2011, pp. 714-730.

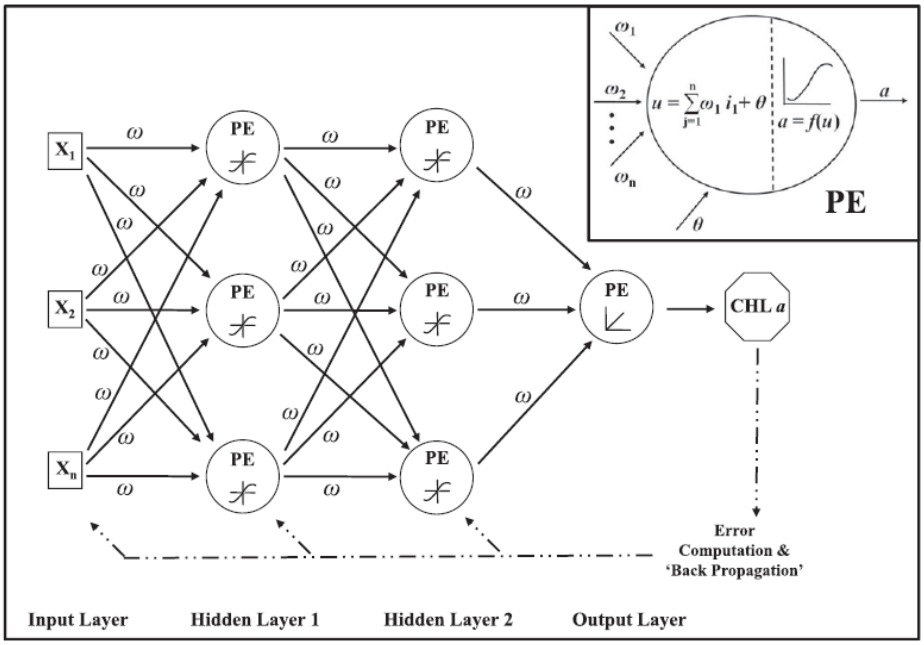

Figure 8. Artificial Neural Network [25].

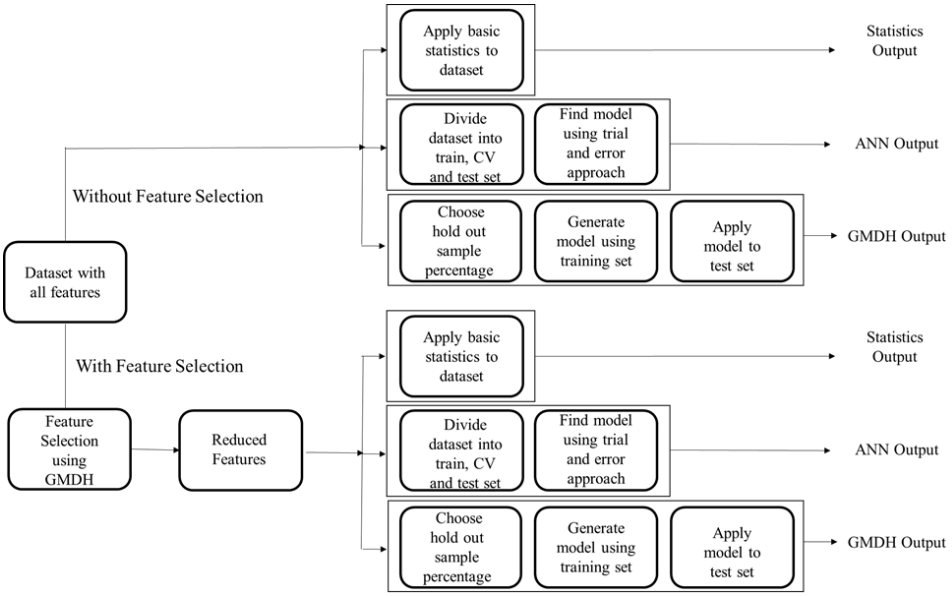

Figure 9. Methodology of Case Study.